2011

Optimal and Heuristic Lead-Time Quotation For an Integrated Steel Mill With a Minimum Batch Size

Susan A. Slotnick
Cleveland State University, s.slotnick@csuohio.edu

Publisher's Statement

NOTICE: this is the author's version of a work that was accepted for publication in European Journal of Operational Research. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in European Journal of Operational Research, 210 (2011), 10.1016/j.ejor.2010.09.031

Original Published Citation

Slotnick, S. A. (2011). "Optimal and Heuristic Lead-Time Quotation For an Integrated Steel Mill With a Minimum Batch Size". European Journal of Operational Research, 210, pp. 527–536.

Optimal and heuristic lead-time quotation for an integrated steel mill with a minimum batch size

Susan A. Slotnick *

Department of Operations and Supply Chain Management, Nance College of Business Administration, Cleveland State University, 2121 Euclid Avenue, Cleveland, OH 44115, USA

1. Introduction

Steel production involves a number of processes, from melting the raw materials, purifying and adding alloys, to casting, rolling, and finishing the product (see Fig. 1). To ensure on-time delivery, the steel producer must take into account processing time, as well as potential delays, from when an order arrives at the facility until the finished material is loaded for transport. The model presented in this paper focuses on the first stage of steel production: melting and casting. At the steel mill that motivates this research, the continuous caster is the bottleneck operation. In order to promise delivery dates that are both attainable and acceptable to customers, the steel producer needs to know the expected time from order entry to delivery, which includes the lead-times (processing time plus delay time) for the production processes. This paper develops a model of lead-time policies for the bottleneck process, the continuous caster, which determine in turn delivery promises to the customer.

The consequences of inaccurate delivery promises are twofold. If the promise is too short, then the order is more likely to be delivered tardy. This may result in penalties, refunds to the customer, and/or lost business in the future. If the promise is too long, the customer may decline it and seek out a competitor who promises earlier delivery. Accurate internal lead-time quotations are necessary to determine external delivery-date promises that are both realistic and acceptable to customers.

The author spent four months interacting with members of a task force charged with improving delivery performance at a steel mill. During busy periods, when the volume of orders exceeds production capacity, the percentage of on-time orders drops below the performance target. On-time delivery is of vital strategic importance for this firm: while some customers are flexible with regard to delivery performance, other customers will not accept late orders. Another indication of the criticality of delivering on time is the fact that the annual bonus for these managers depends in part on the percentage of on-time deliveries. The importance of delivery performance was also reflected in the existence and make-up of this task force, consisting of managers of the major divisions of the plant (steelmaking, rolling mills, transportation, and sales).

One reason for poor delivery performance at this plant is the way in which delivery promises are made. All products are assigned a "standard lead time," that is, customers placing orders are quoted a delivery date that consists of a constant processing time plus a fixed number of weeks. In busy periods, that fixed number of weeks is adjusted upward, in an attempt to account for delays resulting from congestion. While this method crudely reflects the status of the plant, it is adjusted infrequently, and the decision is made at headquarters in another city, rather than locally by managers who know about and are sensitive to the current status of production.

In addition, no adjustment is made for the waiting time that will be incurred by an order that is smaller than the minimum batch size. Because of the size of the ladle in which the steel is

* Corresponding author. Tel.: +1 216 687 3876; fax: +1 216 687 9343.
E-mail address: s.slotnick@csuohio.edu

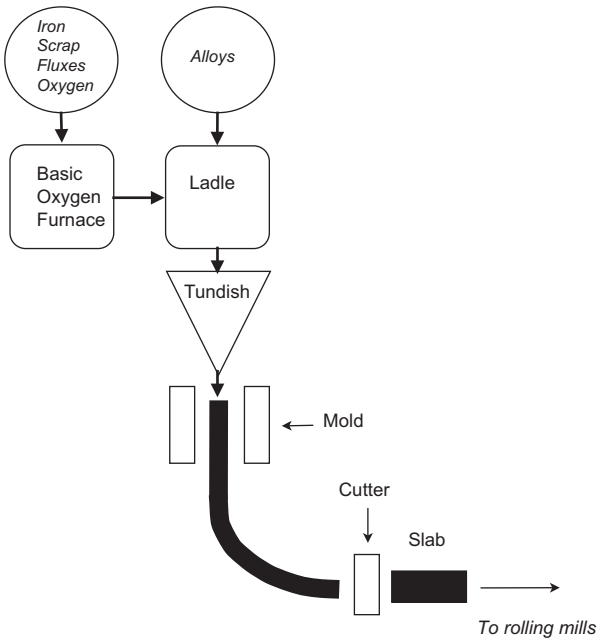

Fig. 1. Steelmaking with continuous casting.

prepared for casting, there is a minimum size for each batch, or melt. Only one product type, or *grade*, can be melted at a time; product differences include the nature of the alloy and the quality of the steel. Since the facility is strictly make to order, producing extra product to stock is not permitted. So customer orders that do not meet the minimum batch size must wait for other orders of the same grade, that is, orders with similar metallurgical and physical characteristics, until there is enough volume to constitute a melt. As a consequence, such orders are more likely to be delivered late.

How should delivery promises be determined in this situation? The current status of the mill, including waiting time for the bottleneck process, is one important component. Orders that are smaller than the minimum batch size may experience an additional delay. So lead-time quotations for the caster, and consequently, customer delivery promises, should take into account processing time, queueing time *and time for arrival of enough tonnage to complete the minimum batch size requirement.*

To provide insights into how this steelmaker (and similar firms) might improve its delivery performance, the present paper develops a model of lead-time policies for the continuous caster, with stochastic arrivals of multiple products, and a minimum batch size. The combination of these elements in one model is a contribution of this paper to the substantial literature on scheduling steel production. The problem is modeled as a stochastic dynamic program with a large state space (two vectors and two scalars). A computational study investigates the relationship of an optimal lead-time quotation to the amount of this product already on order, and to two characteristics of the orders (arrival rate and customer attitude toward delivery promises). For example, will a relatively higher arrival rate of a product, all else equal, result in a longer lead time because of congestion, or a shorter lead time because the minimum batch size is achieved more quickly? What are the effects of arrival rates of other products? It seems intuitive that an order from a customer who is more sensitive to delivery promises, all else equal, should be relatively shorter, but what is the influence of the lead-time sensitivities of other customers?

The results of the computational study provide answers to these questions. First, the optimal lead-time for an order is decreasing in the amount of that product that is already waiting to be melted; this reflects the fact that the time waiting for the orders to accumulate to the minimum batch size will be shorter if there is more there to start with. Second, an order for a product with a higher arrival rate, all else equal, results in a shorter optimal lead-time, since the orders accumulate more quickly and the minimum batch size is reached sooner. Third, higher arrival rates of other products cause more congestion in the system, and so result in a longer optimal lead time for the incoming order. Fourth, optimal lead times are shorter for those customers who are sensitive to the promise date, and might seek other suppliers if the date were too long. Finally, higher sensitivities of other customers result in a shorter optimal lead-time quotation for the current order, since the likelihood of balking is higher, which would reduce congestion and thus shorten waiting-time delays for the current product. These results are intuitive, but not obvious, since the trade-off between longer and shorter delivery promises affects both retention rate (hence incoming revenue) and tardiness (hence discounts or lateness penalties). None of these insights were being used at the steel-making facility.

The near-monotone property of lead times in relation to the level of waiting orders suggests a heuristic approach. A numerical example shows that using this property would save about 40% of computational effort in a typical problem. Computational results confirm that the heuristic is much faster than the optimal procedure, and very close to optimal in value.

The contributions of this research include the analysis of a problem common to steel mills and other applications (such as the production of glass, paint and pharmaceuticals), that is, a lead-time decision that is complicated by a minimum batch size constraint. The computational study provides insights into the problem of how to quote lead-times in order to arrive at delivery promises that will balance the benefits of order retention (revenue) and tardiness costs. The model also provides insights into how aspects of the shop and of customer orders should influence lead-time quotation, and constitutes an application of stochastic dynamic programming to a practical problem of high dimensionality.

The rest of the paper is structured as follows. Section 2 discusses related research in the areas of scheduling steel production and lead-time models. The model is presented in Section 3, properties and algorithms are described in Section 4, and the computational study is detailed in Section 5. Section 6 presents a summary and conclusions.

2. Related work

2.1. Scheduling steel production

The past twenty years have seen scores of papers on the topic of scheduling steel production. Surveys are included in Lee and Murthy (1996), Cowling and Rezig (2000), Dutta and Fourer (2001), Tang et al. (2001), Tang and Wang (2008). Tang et al. (2000) review mathematical programming models and expert-system approaches to scheduling steel production. The present discussion covers only those papers that focus on scheduling continuous casters, with special attention to minimum batch size and related grouping decisions.

Two papers consider the problem of scheduling a caster for which product changeovers involve costly setups. Box and Herbe (1998) present a decision support system for a twin strand continuous caster. A heuristic algorithm selects and then sequences orders for casting, with the objective of minimizing "pseudo-costs," which include penalties for inefficient use of capacity, missing due-dates, and time spent in setups. In order to accommodate minimum batch sizes for the caster, they recommend producing future

as well as present and past-due orders of low-volume cast families, which may also involve making some tonnage to stock. Dorn et al. (1996) apply iterative improvement techniques to scheduling a continuous caster, using constraint satisfaction to minimize setups and product changeovers, as well as to take into account product compatibility and due-dates.

A number of studies develop models for scheduling the caster and subsequent operations. Tamura et al. (1998) employ a two-stage algorithm (including both human and computer components) to schedule casting and hot rolling. Multiple objectives incorporate product selection and due-dates. Constraints take into account chemical and physical characteristics, batch sizes and other capacity constraints. Macro- and micro-level schedules are produced by a combination of human expertise, search and dynamic programming.

Cowling and Rezig (2000) use mathematical programming and heuristics to find near-optimal integrated schedules for a continuous caster and (downstream) hot strip mill. The compound objective includes due-date performance, flow time and set-up considerations, with constraints based on physical dimensions and sequencing and capacity. The three-stage algorithm first finds a feasible solution, next improves it with greedy methods, and finally uses local search for further improvement. Computational tests using industrial data demonstrate the superiority of this approach over manual solutions, and potential cost benefits include savings from reduced inventory and energy, as well as increased flexibility and throughput.

Cowling et al. (2004) employ a multi-agent system to schedule casting and milling operations, in the face of real-time events such as material shortages and rush orders. The scheduling of the continuous caster defines production sequences (heats) under constraints involving physical and chemical compatibility. The model incorporates linear programming and bin-packing, with the objective of minimizing the number of heats and earliness/tardiness penalties. The minimum batch idea is embodied in the scheduling of heats. A simulation study shows that this method dominates a centralized approach in which schedules are generated sequentially without coordination among the different processes.

The problem of grouping orders for a continuous caster involves the minimum batch size of the ladle, the cost of setups and product turnover, and the desire to minimize or eliminate making product to stock. Chang et al. (2000) use integer programming and a column-generation heuristic that minimizes the total number of casts required to group a given set of charges (ladle loads) for the casting operation. Constraints include product characteristics, time limits, and technological constraints.

Ferretti et al. (2006) use an Ant Colony metaheuristic applied to a Traveling Salesman Problem to schedule continuous casting and billet cooling. The objective is to determine the sequence of jobs that maximizes profit (revenue of sold billets minus the costs of billet stocking and order delays). In comparison with a mixed integer linear approach, this algorithm models the industrial situation more closely, and provides good solutions in acceptable computing times.

Tang et al. (2000) develop a four-step scheduling model for steelmaking, refining and continuous casting. In order to ensure production continuity and just-in-time delivery, the model minimizes penalty costs from product changes, temperature drops and earliness/tardiness. Rough scheduling is performed manually, and then the nonlinear model is transformed into a linear programming model which can be solved by standard methods. A subsequent paper (Tang et al., 2002) develops an integer programming model for the same problem, which minimizes "cast breaking, job waiting and earliness/tardiness" and is solved by a combination of Lagrangian relaxation and dynamic programming.

Naphade et al. (2001) consider a problem of melt scheduling that shares some characteristics of continuous casting: a minimum batch size and different families of products. Their method minimizes waste (extra material that is melted to meet the minimum batch size of an order) and total tardiness. A mixed-integer program is computationally expensive, and so a heuristic is formulated that decouples order selection and resource allocation. This method is tested computationally and has been implemented as part of a decision support system.

Dobson and Nambimadon (2001) consider the combination of batching and scheduling that is typical of heat treatment (including casting) in steel production. Jobs belong to different families, which share processing characteristics. Batches cannot be split. The model minimizes weighted flow time, which they describe as a proxy of both cost and lead time. Processing times are not dependent on batch contents, jobs may be of different sizes, and all jobs are available for processing at the time of the decision. They formulate an NP-hard integer program, decompose it into scheduling and batching problems, and develop an optimal polynomial-time algorithm for the special case of related job volumes. Heuristics are developed for the general case.

Tang and Zhao (2008) look at the problem of heating billets as a "semi-continuous" batching machine. That is, a batch consists of orders that share processing characteristics, and orders enter and leave the processor individually, with the processing time of the batch determined by the longest processing time of any one job. The objectives are minimization of makespan and total completion time. Processing times are deterministic. Optimal properties are derived for the batching and scheduling problems, and used to develop dynamic programming algorithms.

Tang and Wang (2008) develop a decision support system to perform two levels of order grouping for a continuous caster. Jobs are grouped into *charges* by grade (chemical composition) and size, and charges are grouped into *casts*, which must consist of compatible grades. The objectives include minimization of "quality upgrading" (delivering a higher quality steel than the customer specified, in order to achieve a minimum batch size), late delivery and extra inventory (for the charge batching problem), and minimization of setups and deviations from priorities and target weights (cast batching problem). A mixed integer-linear program integrates the two problems, and heuristic algorithms obtain solutions of practical problems with acceptable running times. Computational results show that the proposed methods generate solutions that are better and faster than manual methods. Another paper that considers adapting customer specifications, by Balakrishnan and Geunes (2003), presents a production planning model that accounts for the possibility of customer flexibility in product specifications (such as specialty steel).

These articles solve variations of scheduling problems associated with steel production, including caster scheduling, integrated schedules for the caster and downstream operations, batching and order grouping. Batch size considerations are sometimes included as constraints, and/or incorporated into minimization of setup and changeover costs. One way of dealing with the problem of orders that are smaller than minimum batch size is to consider the balance of the batch as make-to-stock product; another is to take advantage of customer flexibility, or combine orders so that some customers receive a higher quality than specified.

The present paper takes a different view of the minimum batch size problem. Instead of seeking ways to use the extra product that would result from rounding up and producing a batch that meets the minimum batch size but is more than has been ordered, or minimizing setup and other costs while incurring penalties for late delivery, the research presented here seeks to maximize revenue (offset by tardiness penalties) by finding lead-time policies that take into account the time needed to accumulate enough orders

for those low-volume jobs that must wait for others until a minimum batch size is achieved and production is initiated. The next subsection places this research in the context of the body of literature that focuses on lead-time quotation.

2.2. Lead-time models

Although the studies of lead-time quotation and due-date setting are intertwined, and "lead-time quotation" is often used synonymously with "due-date setting," the following distinction will be made in this paper. *Lead time* is the processing time plus the amount of time that an order incurs from waiting and other types of delays in the manufacturing facility. *Lead-time quotation*, an internal policy, is necessary for setting *external delivery promises* given to customers. This section focuses on those contributions that are most closely related to the present paper, in particular, research that uses lead time as a decision variable.

The literature on due-date setting in deterministic models is reviewed by Cheng and Gupta (1989). Under the general title of "Due-Date Management Policies," Keskinocak and Tayur (2004) provide a comprehensive survey of the literature on lead-time quotation. Relevant to the present paper is their discussion of due-date management research with order selection decisions, in which demand (the orders retained or refused) is affected by the lead-time policy and subsequent due-date quotation, papers that combine pricing and order selection, and papers that explicitly take into account shop load to quote lead times and set customer due-dates. Missbauer and Uzsoy (2010) discuss workload-dependent lead-time models in the context of production planning optimization. For more recent reviews of this literature, see Zorzini et al. (2008), Upasani and Uzsoy (2008).

Two previous papers consider lead-time quotation using profit maximization and tardiness penalties. In Chatterjee et al. (2002), the sales department knows the processing time of an order, but does not have complete information about the status of the shop or current delays. Customers decide to stay or leave, depending on the quoted delivery promise; the optimal lead-time policy is a log-linear function of the processing time of the order. Slotnick and Sobel (2005) extend this model to the case where the firm has complete information about shop status and delays, and compare the performance of the two procedures to gain insights about when it is cost-effective for a firm to expend resources to provide information about processing backlogs to the sales department.

A number of papers consider lead-time policies when due-dates must be reliable. Kaminsky and Lee (2008) develop due-date quotation algorithms for a dynamic single processor problem with the objective of minimizing the sum of quoted due dates, with no tardiness allowed. Ata and Olsen (2009) investigate the effect of the shape of the customers' delay cost function on capacity, lead-time and sequencing decisions in a make-to-order facility. Late delivery is not allowed, customers leave if the lead-time quotation is too long, and production times are deterministic, as are the delay costs. The model maximizes revenue minus the cost of capacity. Results include asymptotically optimal policies for different delay cost functions, which yield cost savings as demonstrated by a numerical study.

Several recent papers consider the problem of determining due-dates for different customer classes. Plambeck (2004) develops a model for two customer classes, differentiated by price and patience, and reliable due-dates. When a customer arrives, the manager quotes a delivery date and then sequences orders. An upper bound on the long-run average profit rate is derived from a static formulation of the problem, and if some customers will accept long lead times, then capacity utilization will be near 100%. The author uses diffusion approximations to derive a near-optimal policy, with a closed-form expression for profit rate. A simulation study demonstrates the accuracy of these approximations.

Kapuscinski and Tayur (2007) present a finite-horizon, discrete-time model with stochastic demand, deterministic processing time, and two classes of customers with different delivery-time sensitivities and contribution margins, which require reliable due-dates (no late deliveries allowed; early shipments are not penalized). All orders are accepted, and demand and price are not related to due-date quotation. The tradeoff here is that an early delivery quotation to a low-margin customer could cause a future high-margin customer to wait longer. The authors characterize the optimal policy and develop an approximation that provides fast near-optimal solutions.

Celik and Maglaras (2008) consider the tradeoff between price and lead times for multiple customer classes. They also allow dynamic selection of price and lead times, after which orders are sequenced and may be expedited at a cost. With the objective of maximizing the long-run average expected revenue minus the cost of expediting, a diffusion model approximation is developed and evaluated via computational studies. This paper includes the features of expediting capability and the option for a product to be offered at multiple lead times with different prices.

The present paper extends the current literature by developing a stochastic model in which lead-time policies depend on current workload *and* on the time needed to accumulate orders to satisfy a minimum batch size constraint. As in order acceptance models, customers may choose to stay or not, depending on the quoted lead time. The computational study provides insights into the relationship of lead-time policies, number and size of orders already received, customer characteristics (the tolerance of customers for long lead-times) and order characteristics (the arrival rate of a particular class of orders) for the quotation of lead-times in systems that require a minimum batch size for processing.

3. Caster system model

The customer orders in this discrete-time Markov decision process (MDP) model arrive as prospects who are issued lead-time quotations. Period n is the interval of time between the arrivals of the nth and $n+1$st orders, and the planning horizon is the time needed for N prospective orders to arrive. At the beginning of period n, the decision maker learns a_n and g_n which are the respective tonnage and grade of the nth prospective order, $n = 1, 2, \ldots, N$. Then the decision maker issues a lead-time quotation (a decision variable) $L_n + \Theta$ (where L_n is the lead time in the caster and Θ is the processing time on the subsequent operations) to the prospective customer with order n, who reacts by making a firm commitment or by balking. Let δ_n indicate the reaction; $\delta_n = 0$ if the customer balks, and $\delta_n = 1$ if not. Since the caster is the bottleneck operation, the rest of this paper focuses on the decision variable L_n; for purposes of this analysis, Θ is treated as a constant.

In the model, the prospective customer with order n balks with probability $1 - e^{-\xi(g_n)L_n}$ and submits a firm order with probability $e^{-\xi(g_n)L_n}$. That is, the probability of a tentative order becoming firm is a convex decreasing function of the lead-time quotation L_n, and $\xi(g_n)$ parameterizes the "impatience" of the prospective customer with the nth order. In other words, $\xi(g_n)$ parameterizes the sensitivity to the delivery promise of the prospective customer with order n: for a given value of L_n, the higher the value of $\xi(g_n)$, the more likely that the customer will not submit order n. Customer impatience depends on grade, since in this facility, particular grades can be associated with particular customers. For example, important customers with large volumes of orders, such as automotive fabrication plants, demand short lead times and on-time delivery, or they will take their business elsewhere. On the other hand, some

customers place their orders far enough in advance of their production so that they can accept a longer lead-time without disrupting their own production. Customer patience as a decreasing function of lead-time is common in the lead-time literature (Dobson and Pinker, 2006; Ata and Olsen, 2009), and there is also precedent for using the exponential form (Duenyas and Hopp, 1995; Chatterjee et al., 2002; Slotnick and Sobel, 2005).

If the prospective customer places a firm order n, the tonnage of that order is added to a virtual "bucket"; that is, a database entry keeps track of the volume of orders for each grade g (among a total of G grades). If the customer balks, the level of the bucket is unchanged. Whenever a bucket is full, that is, the accumulated volume of orders of that grade has reached the minimum batch size, those orders join the caster queue as a single batch. See Fig. 2. The chemical composition determines which products can be produced in the same melt. As described in Tang and Wang (2008), it is possible to produce more than one type of order in a melt, by delivering a higher-quality product to some customers. Even allowing for such combinations, there are hundreds of possible melt formulas. The decision of how to combine similar grades into melts is not considered in the present paper. So "grade," in this paper, is used to refer to a grade family, that is, a group of products that can be melted together.

Following the arrival of order n, let T_n be the elapsed time until the bucket for grade g_n is full and ready to join the caster queue. Let W_n be the subsequent time that the order spends in the caster system. The order is tardy in the caster system if $T_n + W_n > L_n$. Note that, because the caster is the bottleneck operation, the delivery promise to the customer is a function of the lead-time quotation for the caster system. For ease of exposition, the balance of the paper refers to the influence of lead-time quotations on customers, meaning the influence of delivery promises *derived from* these lead-time quotations. So an order that is tardy in the caster system is assumed to be delivered tardy to the customer.

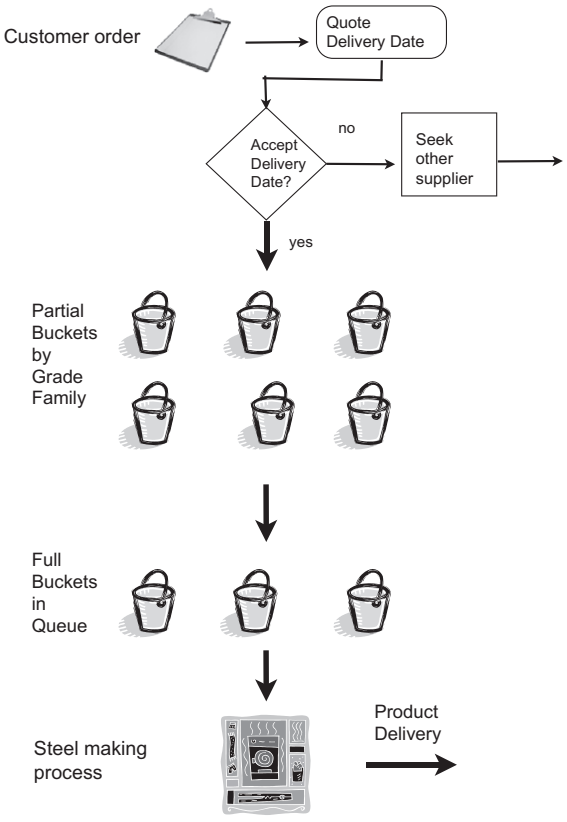

Fig. 2. Order flow and buckets.

If the prospective customer places a firm order n and the order is *not* tardy, then the mill is credited with $w^r(g_n)a_n$ which is proportional to the tonnage of the order (a_n). The profit rate $w^r(g_n)$ depends on the grade because different grades are priced differently. If the order *is* tardy, then the credit is offset by a penalty w^p; so the net credit is $w^r(g_n)a_n - w^p$. Let τ_n indicate whether this order n is tardy or not; that is, $\tau_n = 1$ if the order is tardy, and $\tau_n = 0$ if not. The factor w^r represents the value to the firm of processing and delivering an order; it may be considered a per-unit revenue or contribution margin.

With this notation, the total credit that the mill accrues during the planning horizon is

$$\sum_{n=1}^{N} \delta_n [w^r(g_n)a_n - \tau_n w^p].\psi \qquad (1)$$

After the arrival of order n, the steel mill managers already know the grade g_n and tonnage a_n of the prospective order, as well as the lead-time quotation L_n which they have chosen. So there are three sources of randomness. First, the prospective customer may balk. Second, if the customer does not balk, the interval of time until the bucket for grade g_n becomes full is random. Third, the subsequent time spent in the caster system is random. Thus, after the mill issues quotation L_n, the expected value of its net credit associated with prospective order n is

$$e^{-\xi(g_n)L_n}[w^r(g_n)a_n - w^p P\{T_n + W_n > L_n\}].\psi \qquad (2)$$

Therefore, the expected value of the total credit during the planning horizon is

$$E\left(\sum_{n=1}^{N} e^{-\xi(g_n)L_n}[w^r(g_n)a_n - w^p P\{T_n + W_n > L_n\}]\right).\psi \qquad (3)$$

Expression (2) is the expected single-period reward in the MDP model, and the overall criterion is to maximize the expected total credit, i.e., (3), with respect to non-anticipative decision rules for choosing lead-time quotations $L_1, L_2, \ldots, L_N$. The analysis is based on the following assumptions regarding the stochastic elements in the model. The tonnages and grades, $(a_1,g_1),(a_2,g_2),\ldots,(a_N,g_N)$, are independent and identically distributed pairs. That is, they have an exogenous joint distribution which does not depend on lead-time quotations directly or indirectly. The arrival times of prospective orders of various steel grades are independent grade-dependent Poisson processes. The tonnages of different orders with the same grade are independent and identically distributed exponential random variables. The arrival rate of each grade, denoted λ_g for grade g, and mean tonnage per order, are exogenous known parameters.

The times that firm orders spend in the caster system (W_n if customer n does not balk) are independent and identically distributed as the equilibrium waiting time in an M/M/1 queueing system with arrival rate Λ and service rate μ. That is, W_n is exponentially distributed with rate $\Lambda - \mu$. An unavoidable source of complexity in this model is that Λ is endogenous. The rate at which full buckets enter the caster process depends on the exogenous arrival processes *and* on the lead-time quotations. Short quotations will induce infrequent balking and, therefore, the buckets will fill quickly and the input process to the caster will have a higher rate. On the other hand, if most of the quotations are long, then many customers will balk and so the buckets will fill slowly, and the input to the caster will have a lower rate. Congestion will be higher in the former case than the latter, and the W_n's will be stochastically greater. The input rate to the caster in this model reflects these considerations and, when the nth prospective order arrives, is

$$\Lambda = \sum_{g=1}^{G} \lambda_g e^{-\xi(g)\mathcal{L}_g}, \psi$$

where $\mathcal{L}_g$ denotes the most recent quotation on an order of grade g. The idea is that customers are aware of the most recent lead-time quotation through the delivery promise, which affects their decisions to balk or stay.

Given a value of L_n, these assumptions and a normal approximation yield an easily computed value of $P\{T_n + W_n > L_n\}$ (Sobel, 2009) which is needed to optimize the expected value of the total credit (3). This yields a discrete-time Markov decision process (MDP) with the objective of maximizing (3). The action of the MDP is the quotation L and its state is the vector $(\mathbf{s}, a, \boldsymbol{\mathcal{L}}, g)$ with $2G + 2$ components having the following interpretation. At the beginning of period n, i.e., when the nth prospective order arrives, the vector $\mathbf{s} = (s_1, s_2, \ldots, s_G)$ specifies the bucket levels of the grades, a and g are the tonnage and grade of the prospective order, and the vector $\boldsymbol{\mathcal{L}} = (\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_G)$ specifies the most recent previous lead-time quotation for each grade.

In order to specify the dynamics of the MDP, let $(\mathbf{s}, a, L, g)$ be the state upon the arrival of the nth prospective order, and let L be the ensuing lead-time quotation. Let $(\hat{\mathbf{s}}, a', \widehat{\boldsymbol{\mathcal{L}}}, g')$ denote the state when the $(n+1)$st prospective order arrives. The joint probability distribution of (a', g'), the tonnage and grade of that order, is exogenous, i.e., it does not depend either on the state or the lead-time quotation. Also, $\widehat{\mathcal{L}}_g = L$ and $\widehat{\mathcal{L}}_k = \mathcal{L}_k$ for all $k \neq g$ because only the nth prospective order receives a lead-time quotation during period n. Similarly, $\hat{s}_k = s_k$ for all $k \neq g$ because only the grade g bucket is affected by the arrival of prospective order n and its decision to balk or not. If the customer balks, and that occurs with probability $1 - e^{-\xi(g)L}$, then $\hat{s}_g = s_g$.

Let M be the capacity of the ladle, i.e., the minimum batch size. If the nth prospective customer does not balk, there are two cases. If $s_g + a \geqslant M$, then the updated grade g bucket has reached the minimum batch size, the grade g batch enters the caster process, and $\hat{s}_g = 0$ (the bucket is emptied). If $s_g + a < M$, then the updated grade g bucket has not yet reached the minimum batch size and cannot yet enter the caster process, and $\hat{s}_g = s_g + a$. If the nth prospective customer balks, then the amount in the grade g bucket remains unchanged so $\hat{s}_g = s_g$. Since the order volumes of the products in question are relatively small, the model does not keep track of the small amount by which $s_g + a$ may exceed M.

The MDP model corresponds to the following dynamic program with $f_{N+1}(\cdot,\cdot,\cdot,\cdot) \equiv 0$:

$$\begin{aligned} f_n(\mathbf{s}, a, \boldsymbol{\mathcal{L}}, g) = \max_{L>0} \Big\{ & e^{-\xi(g)L}\left[w_g^r a - w^p P\{T + W > L\}\right] \\ & + e^{-\xi(g')L}\left[E[f_{n+1}(\hat{\mathbf{s}}, a', \widehat{\boldsymbol{\mathcal{L}}}, g')\right] \\ & + \left(1 - e^{-\xi(g')L}\right) E\left[f_{n+1}(\mathbf{s}, a', \widehat{\boldsymbol{\mathcal{L}}}, g')\right] \Big\} \cdot \psi \end{aligned} \qquad (4)$$

4. Properties and algorithms

4.1. Introduction

The problem of determining lead-time policies for the steel mill exemplifies the strategic importance of operational decisions, since delivery performance is a vital competitive priority. How can the caster system model (4) contribute to understanding how to quote lead-times? What characteristics of the production facility, customers and orders should the delivery performance team consider when deciding how to determine due-date promises? What are the tradeoffs, and how should they inform decision making?

Answers to these questions are found in the relationship between state variables, model parameters, and the decision variable. It seems reasonable that the combination of current order size a_n and bucket level for the current grade s_g, that is, the updated bucket level $\hat{s}_g$, should influence the estimate of how long it will take to finish and deliver the order, since waiting time until release into the caster system depends on when the bucket level reaches minimum batch size (see Fig. 2).

Other parameters of the model also influence the lead-time decision, which influences in turn the delivery promise to the customer. Exogenous arrival rates λ_g have a dual effect: higher arrival rates may cause congestion delays, but may also speed the filling of the bucket which shortens the wait until minimum batch size is reached. If customers vary in their sensitivity to the length of due-date promises, it seems likely that more impatient customers (higher $\xi(g)$) should be quoted shorter delivery promises. The sensitivity of other customers may have two countervailing effects. If resources are devoted to meeting the higher impatience of others, a longer due date should be quoted for the current order. On the other hand, the impatience of others may lead to fewer retained orders, lower congestion and so the ability to quote a shorter delivery date for the current model.

The richness of the model (4) makes possible the investigation of these questions. However, the resulting complexity brings difficulties to the analysis. The size of the state space, which includes two vectors and two scalars, invokes the "curse of dimensionality." The next subsection describes how a near-monotonone property of the model provides an opportunity for a fast, high-quality heuristic.

4.2. Computational acceleration

The number of arguments in the dynamic program (4) and the well-known curse of dimensionality make the computation of an optimal solution a formidable task. However, if the optimal lead-time quotation were non-increasing in the size of the updated bucket $\hat{s}_g$, there would be a dramatic reduction in computational effort. First, consider the number of states that are visited in a straightforward implementation of the dynamic program. Let M be the minimum batch size in units (tons), G be the total number of possible grades, K be the largest acceptable lead time, and A be the largest order size. Then there are G components in each of the vectors $\mathbf{s}$ and $\boldsymbol{\mathcal{L}}$. The state space has $(MK)^G A\ G$ elements. So for N arrivals, making comparisons for K possible values of L, the total number of comparisons is $(NK)(MK)^G(AG)$.

The near-monotonicity with respect to $\hat{s}_g$ can be exploited as follows. At each stage n of computation, and for each state vector, begin by performing K comparisons to calculate the optimal lead time $\widetilde{L}_0$ with an arrival to an empty bucket ($s_g = 0$). So if $K = 5$, then there would be five comparisons. For the next value of $\hat{s}_g$, i.e., $\hat{s}_g = 1$, the values of L to be considered are $1, 2, \ldots, \widetilde{L}_0$. If $\widetilde{L}_0 = 3$, only three values of L need be considered for the state in which $\hat{s}_g = 1$.

To illustrate the savings in effort, let $M = 5$, $K = 5$, $A = 5$ and $G = 3$. The straightforward approach would require $1{,}171{,}875N$ comparisons. If $\widetilde{L}_1 = 4$ and $\widetilde{L}_2 = 2$, using the monotonicity results would decrease this to $703{,}125N$ comparisons, which is about a 40% saving. Computational experiments show that the lead-time policy is not always non-increasing in $\hat{s}_g$, so this property is employed as a heuristic in the computational study described below.

4.3. The algorithms

The computational study described in Section 5 employs two solution procedures: an implementation of the dynamic program (4) (OPTDP), and a heuristic based on the near-monotone property described above (MONDP). OPTDP loops through the grid defined by the state variables (grade, order size, vector of previous

lead-time quotations, vector of bucket levels) and saves the lead-time quotation that provides the highest profit for that state. The implementation is straightforward:

Procedure OPTDP

For each grade;
For each order size;
For each value of previous lead-time quotation;
For each bucket level;
For each possible lead-time quotation;
{Calculate the probability of tardiness;
Calculate the expected profit
If the current profit is the highest, save this lead-time quotation and profit}

The probability of tardiness is calculated using the following formula (Sobel, 2009):

$$P\{T+W>L\}=\sum_{j=0}^{L}\phi\left(\frac{j-\mu_T}{\sigma_T}\right)e^{-(\mu-\Lambda)(L-j)}+\left(1-\Phi\left(\frac{L-\mu_T}{\sigma_T}\right)\right). \quad (5)$$

The mean and standard deviation of time remaining to fill this bucket are

$$\mu_T=\frac{\sqrt{M-(s_g+a)}}{\lambda_g},\quad \sigma_T=\frac{\mu_T}{\lambda_g},$$

where μ is the processing rate of the caster system, Λ is the sum of filtered arrival rates of all grades, and ϕ and Φ are the density function and distribution function of the standard normal distribution.

The second procedure, the heuristic MONDP, uses the near-monotone property of the model described in Section 4.2 to decrease the search space:

Procedure MONDP

For each grade;
For each order size;
For each value of previous lead-time quotation;
For each bucket level;
For each possible lead-time quotation,
where possible lead-times for this bucket are bounded above by the highest lead-time value for the previous bucket in this state;
{Calculate the probability of tardiness;
Calculate the expected profit
If the current profit is the highest, save this lead-time quotation and profit}

5. Computational study

5.1. Design of the computational study

The computational study has two aims. The first is to investigate the relationship between the optimal lead time L and three other values: the level of the updated bucket of orders already received for the current grade $\hat{s}_g$, the exogenous arrival rate λ_g, and customer impatience $\xi(g)$. The influence of a higher λ_g is ambiguous: orders with a higher arrival rate fill the bucket faster, and so the waiting time until joining the queue (T) should be lower. However, a higher arrival rate means potentially more congestion in the shop, a longer queue, and so a longer wait in the queue before processing on the caster (W). Similarly, it seems reasonable to quote a shorter delivery promise to an impatient customer (higher $\xi(g)$), but what is the effect of the delivery-time sensitivities of other customers?

In order to investigate the influence of these three factors on the optimal lead-time quotation, two values each of λ_g and $\xi(g)$ are varied among three grades (see Table 1). Preliminary pilot studies determined that λ_g be set at a high value of 1.3 and a low value of 0.5, and $\xi(g)$ be set at a high value of 0.5 and a low value of 0.005 (these values were chosen to generate a variety of lead-times over all states, rather than values that were all on the boundaries, i.e., 1 and 8). The computational study uses a full-factorial design, resulting in twenty separate tests. In each test, the maximum lead time is set at 8 and minimum batch size is 5. The per-unit reward is 2 and the per-unit time penalty is 1 for all grades. The processing rate of the caster, μ, is set at 4. This design results in 960,000 states per test, or 19,200,000 states for all twenty tests. Each test runs for 20 periods, i.e., 20 arriving orders.

The second aim is to evaluate the speed and quality of the heuristic, as compared to the optimal dynamic programming algorithm. To do this, both OPTDP and MONDP were run on the 20 tests described above, and compared in terms of solution quality and running time. Both programs are coded in FORTRAN 90 and run on a Sun Ultra workstation under the Solaris 8 operating system.

Computational effort in this program is linear in the planning horizon N. So a smaller N means faster processing time, or more computational resources available to expand the problem (for example, to include more grades). To determine whether a horizon of less than $N = 20$ would provide acceptable results, convergence of the lead-time policy is analyzed by comparing differences in lead times and values between the same states in different periods.

Table 2 displays the results of these comparisons for both algorithms. For OPTDP, in five of twenty tests, there is immediate lead-time convergence (no difference between lead-time quotations, for corresponding states, in periods 1 and 20). That is, if computation had been done with $N = 1$ rather than $N = 20$ periods, the optimal lead-time policy would have been the same. Test 2 converged after the second period. In the remaining fourteen tests, the policies in periods 1 and 2 were still different, but the frequency of such differences was very low. The worst case had complete agreement in all but 293 states, that is, in 99.69% out of 960,000 states (see the second and third columns of Table 2, which lists the number and percentage of corresponding states with different lead times in periods 1 and 2, respectively). The convergence with regard to the value function was checked using maximum absolute difference MAXABS (where $\gamma=(\mathbf{s},a,\mathcal{L},g)$):

$$MAXABS=max_\gamma|f_3(\gamma)-f_4(\gamma)|-max_\gamma|f_1(\gamma)-f_2(\gamma)|.$$

The largest value of MAXABS was 0.0003357 (see the fourth column of Table 2).

For MONDP, there are five tests with immediate convergence of lead times. Tests 2 and 7 converge after period four. In the remaining thirteen tests, the worst case had complete agreement in 97.74% out of 960,000 states (see the fourth and fifth columns of Table 2). The largest value of MAXABS was 0.8096924 (see the sixth column of Table 2). These results demonstrate that both procedures converge rapidly, with small values of N giving excellent results most of the time.

Table 1
Parameter settings for the computational study.

Parameter	Test number									
By grade	1	2	3	4	5	6	7	8	9	10
λ_1	1.3	1.3	0.5	0.5	1.3	0.5	0.5	1.3	1.3	0.5
λ_2	1.3	1.3	0.5	0.5	0.5	0.5	0.5	0.5	1.3	0.5
λ_3	1.3	1.3	0.5	0.5	1.3	1.3	1.3	1.3	1.3	0.5
$\xi(1)$	0.5	0.005	0.5	0.005	0.5	0.5	0.005	0.005	0.5	0.5
$\xi(2)$	0.5	0.005	0.5	0.005	0.5	0.5	0.005	0.005	0.5	0.5
$\xi(3)$	0.5	0.005	0.5	0.005	0.5	0.5	0.005	0.005	0.005	0.005
	11	12	13	14	15	16	17	18	19	20
λ_1	1.3	0.5	1.3	0.5	0.5	1.3	1.3	1.3	1.3	1.3
λ_2	1.3	0.5	0.5	1.3	0.5	1.3	0.5	1.3	0.5	0.5
λ_3	1.3	0.5	1.3	1.3	1.3	0.5	0.5	0.5	0.5	0.5
$\xi(1)$	0.5	0.5	0.5	0.5	0.005	0.5	0.005	0.5	0.005	0.5
$\xi(2)$	0.005	0.005	0.005	0.005	0.005	0.5	0.5	0.005	0.5	0.5
$\xi(3)$	0.005	0.005	0.005	0.005	0.5	0.005	0.5	0.5	0.005	0.005

Table 2
Convergence for the computational study.

Test	OPTDP			MONDP		
	# Diff. states	% Diff. States	MAXABS	#Diff. states	% Diff. states	MAXABS
1	0		0.0003357	0		0.0003357
2	0		0.0001526	0		0.0001526
3	0		0.0002441	0		0.0002441
4	0		0.0000763	0		0.0001221
5	0		0.0003128	0		0.0003128
6	0		0.0002785	0		0.0002785
7	242	0.025	0.0000763	0		0.0000916
8	244	0.025	0.0000763	249	0.026	0.0782318
9	60	0.006	0.0002975	3050	0.318	0.8096924
10	54	0.006	0.0002136	126	0.013	0.0713806
11	293	0.031	0.0001602	3855	0.402	0.3947678
12	225	0.023	0.0001526	59	0.006	0.0017014
13	134	0.014	0.0001678	12412	1.293	0.8095169
14	214	0.022	0.0001144	21686	2.259	0.7382126
15	21	0.002	0.0001907	1	0.000	0.0003510
16	74	0.008	0.0002823	49	0.005	0.0676575
17	52	0.005	0.0002060	45	0.005	0.0006027
18	49	0.005	0.0002441	62	0.006	0.8078232
19	131	0.014	0.0001526	1022	0.106	0.7775955
20	117	0.012	0.0002441	15	0.002	0.0705185

5.2. Results of the computational study

The relationship of optimal lead-time quotation to updated bucket level, customer impatience and arrival rate was analyzed using regression analysis with 19,200,000 observations (the output of OPTDP). The regression model is:

$$L = \beta_0 + \beta_1 \hat{s}_g + \beta_2 \xi(g) + \beta_3 \lambda_g + \beta_4 INT1 + \beta_5 INT2 + \epsilon.\psi$$

Two interaction terms are included, to assess the relationship between customer impatience and arrival rate of the current arrival and these two characteristics of the other two grades. The first interaction term (*INT*1) represents the relationship between the impatience of the customer of the current order (grade g) and the impatience of other customers, and the second interaction term (*INT*2) represents the relationship between the arrival rate of the grade of the current order and the arrival rates of other grades:

$$INT1 = \xi(g) \sum_{j \neq g}^{G} \xi(j) \leftarrow INT2 = \lambda_g \sum_{j \neq g}^{G} \lambda_j.\psi$$

The regression was run using the REG procedure of SAS (version 9.13). The combination of relatively high R^2 (0.59) and low p-values suggest that the model is a good fit. All variance inflation factors are less than five, so multicollinearity is not a problem here. See Table 3. A plot of the residuals shows that the assumption of normality is justified.

These results provide insights into the effects of customer/order characteristics on lead-time quotation, and answer the questions posed in Section 4.1. The main effects are as follows. Lead-time quotations should be *shorter* when the level of the updated bucket $\hat{s}_g$ is *higher*, and when customers are more impatient (higher $\xi(g)$). The first result confirms that the model reflects the reality of the motivating phenomenon. The second main effect answers one question about the countervailing effect of arrival rates: optimal lead times are *shorter* when the arrival rate of the current grade (λ_g) is *higher*; the bucket fills up faster, and so waiting time to join the queue is lower.

Table 3
Results of regression analysis.

Variable	β	t-Value	p-Value	Variance inflation
Intercept	7.05634	4779.58	<0.0001	0
$\hat{s}_g$	−0.40881	−2235.7	<0.0001	1.00000
$\xi(g)$	−6.77428	−3013.7	<0.0001	2.31438
λ_g	−1.44011	−887.22	<0.0001	3.15211
INT1 (ξ)	−0.18204	−65.12	<0.0001	2.31663
INT2 (λ)	0.67382	1085.18	<0.0001	3.15436

Table 4
Comparison of OPTDP and MONDP (running time in seconds).

Test	1	2	3	4	5	6	7	8	9	10
Avg % Dev	0.0000	0.0003	0.0000	0.000	0.000	0.000	0.667	0.185	0.276	0.561
Max % Dev	0.0000	0.0003	0.0000	0.000	0.000	0.000	1.017	0.349	1.020	0.876
Opt time	2338	2317	2332	2309	2336	2334	2308	2310	2329	2327
Heur time	535	2277	534	1489	536	535	1435	1529	662	726
% Opt time	22.88	98.27	22.90	64.49	22.95	22.92	62.18	66.19	28.42	31.20
	11	12	13	14	15	16	17	18	19	20
Avg % Dev	0.073	0.661	0.497	0.199	0.258	0.549	0.107	0.179	0.544	0.564
Max % Dev	0.401	1.053	1.125	0.827	0.637	0.868	0.762	0.979	1.161	0.881
Opt time	2319	2321	2324	2321	2325	2335	2328	2329	2334	2339
Heur time	1036	1088	1030	971	1276	738	729	718	956	734
% Opt time	44.67	46.88	44.32	41.84	54.88	31.61	31.31	30.83	40.96	31.38

The interaction effects provide additional insights into the tradeoffs between impatience levels and arrival rates of the current order versus other orders. As discussed in Section 4.1, it is not obvious whether higher impatience of other customers will delay the current order, resulting in a longer lead-time, or reduce congestion because impatient customers are more likely to leave, resulting in a shorter lead time. The negative coefficient of INT1 suggests that, given the impatience of the current customer, *higher* impatience levels of other customers result in a *shorter* lead time for the current order. Intuitively, if other customers are more impatient, there is likely to be more balking, and so less congestion at the caster queue. This order will experience less delay, and so it makes sense to quote a shorter optimal lead-time.

The positive coefficient of INT2 suggests that *higher* arrival rates of other grades (given an arrival rate for the grade of the current order) result in a *longer* optimal lead time for the current order. This is because higher arrival rates of other grades will cause more congestion at the caster queue, and more tardiness, if lead-time estimates are not adjusted upward.

These results are intuitively consistent with what might be expected, but are not obvious from the industrial situation or the model itself. In particular, the delivery promises should take into account not just the general congestion of the facility, but also the waiting time that is incurred by orders that are less than minimum batch size. In addition, they provide a range of easily observable measures by which to adjust lead-time quotations: the current volume of orders for each grade; history of order volume across grades; and the experience and opinion of the sales department about the time-related tolerance of various customers. For the steel producer that motivated this study, it was common knowledge on the shop floor which customers had low tolerance for long delivery promises.

The comparison of OPTDP and MONDP reveals that the heuristic produces results very close to optimal, with much faster running times. The average deviation in value function was always less than 1%, and maximum deviation in value function was always less than 2%, with the highest average deviation less than 0.7% and the highest maximum deviation less than 1.2%. Except for one test for which the heuristic took almost as much time as the optimal procedure, MONDP ran substantially faster, with more than 50% saving in computation time for fifteen of the twenty tests. See Table 4 for details.

6. Summary and conclusions

The contributions of this paper are (a) a dynamic programming formulation for the problem of computing an optimal lead-time quotation when there is a minimum batch size, (b) a heuristic algorithm that performs very close to optimal with considerable saving in computation time, and (c) a computational study that provides insights into the relationship between optimal lead-time quotation and bucket level, sensitivity to delivery promise times and arrival rate.

The consequences of these results have policy implications for the steel-producing facility. Since the optimal lead-time quotation for the caster system is generally decreasing in the updated bucket level for that grade, delivery performance and retention rate would be improved by using information about the state of previously accepted orders, in combination with the size of the incoming order, to determine delivery date promises. Shorter delivery promises should be assigned to orders for which the customer is unlikely to place an order if s/he will wait too long for delivery, and also when other customers are relatively more impatient. Higher arrival rates for the current order will result in shorter lead times, and so delivery promises should be adjusted accordingly. However, higher arrival rates of other customers require a longer lead time, because of added congestion in the caster system.

One motivation for this project was the development by the steel mill of an automated order entry system to be used by their customers. A customer would enter the details of the tentative order, and the system would respond with information including proposed delivery time. If the customer decided to place the order, the relevant information would be sent to the operational database. The research described in this paper could serve as the intelligent engine that generates the delivery times, based on the lead-time quotation model described here. Alternatively, the lead-time quotation program could be implemented as a stand-alone decision tool for the sales force, enabling them to tailor their delivery-time promises in a more sophisticated manner than they have been doing (Section 1).

For input to the order-entry system, arrival rates could be determined from historical demand data (by grade), maximum lead time would be set according to company policy and market standards, and the minimum batch size is determined by the size of the physical ladle. Per-unit reward would be the contribution margin of each product, and per-unit tardiness penalty could be quantified from historical data on expediting costs and actual tardiness discounts. The processing rate of the caster is a known operational factor.

In order to quantify customer impatience, the sales department of the steel mill, who are familiar with their customers over time, might be consulted to rank customers in terms of their sensitivity to lead-time. These rankings could then be used to set initial values of the impatience parameter for a series of test runs of the system, which could subsequently be checked and recalibrated if necessary.

In principle, the present program would need to be scaled up to handle hundreds of grades. However, in practice the number of grades to be processed at any one time would be smaller, because

(1) a limited number of grades are ordered each day, and (2) orders that exceed the minimum batch size are placed immediately in the caster queue, since there is no need to wait for a bucket to fill. The program could easily be modified to "streamline" such orders, providing lead-time quotations via the embedded M/M/1 model of the caster system, while including those arrivals in the calculation of the equilibrium waiting time. If the resulting complexity still led to unacceptably long running times for an interactive system, a batch run could be made periodically (each morning perhaps) to provide "typical" lead times for various products, based on the current status of the caster and the orders waiting to be processed.

References

Ata, B., Olsen, T.L., 2009. Near-optimal dynamic lead-time quotation and scheduling under convex–concave customer delay costs. Operations Research 57 (3), 753–768.

Balakrishnan, A., Geunes, J., 2003. Production planning with flexible product specification: An application to specialty steel manufacturing. Operations Research 51 (1), 94–112.

Box, R., Herbe Jr., D., 1998. A scheduling model for LTV Steel's Cleveland Works' twin strand continuous slab caster. Interfaces 18 (1), 42–56.

Celik, S., Maglaras, C., 2008. Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. Management Science 54 (6), 1132–1146.

Chang, S., Chang, M., Hong, Y., 2000. A lot-grouping algorithm for a continuous slab caster in an integrated steel mill. Production Planning and Control 11 (4), 363–368.

Chatterjee, S., Slotnick, S.A., Sobel, M.J., 2002. Delivery guarantees and the interdependence of marketing and operations. Production and Operations Management 11 (3), 393–410.

Cheng, T., Gupta, M., 1989. Survey of scheduling research involving due date determination decisions. European Journal of Operational Research 38 (2), 156–166.

Cowling, P., Ouelhadj, D., Petrovich, S., 2004. Dynamic scheduling of steel casting and milling using multi-agents. Production Planning and Control 15 (2), 178–188.

Cowling, P., Rezig, W., 2000. Integration of continuous caster and hot strip mill planning for steel production. Journal of Scheduling 3, 185–208.

Dobson, G., Nambimadon, R.S., 2001. The batch loading and scheduling problem. Operations Research 49 (1), 52–65.

Dobson, G., Pinker, E.J., 2006. The value of sharing lead time information. IIE Transactions 38, 171–183.

Dorn, J., Gersch, M., Skele, G., Slany, W., 1996. Comparison of iterative improvement techniques for schedule optimization. European Journal of Operational Research 94, 349–361.

Duenyas, I., Hopp, W., 1995. Quoting customer lead times. Management Science 41 (1), 43–57.

Dutta, G., Fourer, R., 2001. A survey of mathematical programming applications in integrated steel plants. Manufacturing and Service Operations Management 3 (4), 387–400.

Ferretti, I., Zanone, S., Zavanella, L., 2006. Production-inventory scheduling using Ant System metaheuristic. International Journal of Production Economics 105, 317–326.

Kaminsky, P., Lee, Z.-H., 2008. Effective on-line algorithms for reliable due date quotation and large-scale scheduling. Journal of Scheduling 11 (3), 187–205.

Kapuscinski, R., Tayur, S., 2007. Reliable due-date setting in a capacitated MTO system with two customer classes. Operations Research 55 (1), 56–74.

Keskinocak, P., Tayur, S., 2004. Due date management policies. In: Simchi-Levi, D., Wu, S., Shen, Z.-J. (Eds.), Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era. Kluwer, Boston, pp. 485–556.

Lee, H.-S., Murthy, S., 1996. Primary production scheduling at steelmaking industries. IBM Journal of Research and Development 40 (2), 231–253.

Missbauer, H., Uzsoy, R., 2010. Optimization models for production planning problems. In: Kempf, K., Keskinocak, P., Uzsoy, R. (Eds.), Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook. Springer, Norwell, MA.

Naphade, K.S., Wu, S.D., Storer, R.H., Doshi, B.J., 2001. Melt scheduling to trade off material waste and shipping performance. Operations Research 49 (5), 629–645.

Plambeck, E.L., 2004. Optimal leadtime differentiation via diffusion approximations. Operations Research 52 (2), 213–228.

Slotnick, S.A., Sobel, M.J., 2005. Manufacturing lead-time rules: Customer retention vs tardiness costs. European Journal of Operational Research 163 (3), 825–856.

Sobel, M.J., 2009. Tardiness risk at a steel mill. Technical Report TM-824, Department of Operations, Case Western Reserve University, Cleveland, OH.

Tamura, R., Nagai, M., Nakagawa, Y., Tanizaki, T., Nakjima, H., 1998. Synchronized scheduling method in manufacturing steel sheets. International Transactions in Operational Research 5 (3), 189–199.

Tang, L., Liu, J., Rong, A., Yang, Z., 2000. A mathematical programming model for scheduling steelmaking-continuous casting production. European Journal of Operational Research 120, 423–435.

Tang, L., Liu, J., Rong, A., Yang, Z., 2001. A review of planning and scheduling systems and methods for integrated steel production. European Journal of Operational Research 133, 1–20.

Tang, L., Luh, J., Liu, J., Fang, L., 2002. Steel-making process scheduling using Lagrangian relaxation. International Journal of Production Research 40 (1), 55–70.

Tang, L., Wang, G., 2008. Decision support systems for the batching problems of steelmaking and continuous-casting production. Omega 36 (6), 976–991.

Tang, L., Zhao, Y., 2008. Scheduling a semi-continuous batching machine. Omega 36, 992–1004.

Upasani, A., Uzsoy, R., 2008. Incorporating manufacturing lead times in joint production-marketing models: A review and some future directions. Annals of Operations Research 161, 171–188.

Zorzini, M., Corti, D., Pozzetti, A., 2008. Due date (DD) quotation and capacity planning in make-to-order companies: Results from an empirical analysis. International Journal of Production Economics 112 (2), 919–933.